DOCUMENT RESUME

ED 177 197 TM 009 690

AUTHOR Patience, Wayne M.; Reckase, Mark D.
TITLE Operational Characteristics of a Rasch Model Tailored Testing Procedure when Program Parameters and Item Pool Attributes are Varied.
INSTITUTION Missouri Univ., Columbia.
SPONS AGENCY Office of Naval Research, Washington, D.C. Personnel and Training Branch.
PUB DATE Apr 79
CONTRACT N00014-77-C-0097
NOTE 38p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, California, April 9-11, 1979)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Computer Programs; Educational Testing; *Item Analysis; Item Banks; *Mathematical Models; Simulation; Standard Error of Measurement; *Statistical Bias; Test Construction; Test Items; *Test Reliability
IDENTIFIERS Computer Assisted Testing; Latent Trait Model; One Parameter Model; *Tailored Testing

ABSTRACT

Simulated tailored tests were used to investigate the relationships between characteristics of the item pool and the computer program, and the reliability and bias of the resulting ability estimates. The computer program was varied to provide for various step sizes (differences in difficulty between successive steps) and different acceptance ranges (precision with which selected items must match desired item difficulties). Item analysis and ability estimates were based on a one-parameter logistic (Rasch) model. Item pools varied in size from 9 to 181 items, and in shape to include normal, bimodal, rectangular, and postively skewed distributions of item difficulty. In addition, item difficulty distributions were designed to fit either precise theoretical distributions or actual test distributions. The acceptance range parameter was important because it influenced the point at which testing was discontinued; there was a fixed maximum of 20 items but a minimum in some circumstances of 4 items. Two different tailored test programs were used, TREE1P and SIM1P. Results were based on 25 simulated administrations under each of the various conditions. (CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Wayne M. Patience
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED177197

TM009 690

Operational Characteristics of a Rasch Model Tailored Testing Procedure when Program Parameters and Item Pool Attributes are Varied

Wayne M. Patience
Mark D. Reckase
University of Missouri-Columbia

INTRODUCTION

Tailored testing, the selection and scoring of test items administered in an interactive fashion to individual examinees has, within the past decade, become the spearhead for application of latent trait models to achievement and ability measurement. The availability of improved computer technology has contributed greatly to the increase in the number of systems presently in operation which administer tailored or adaptive tests. It should be noted that tailored testing as presented here is synonomous with many other assigned names such as adaptive testing, response contingent testing, or sequential testing. While numerous articles have appeared in the literature which describe the one-parameter logistic model and its application in a tailored testing setting (see for example Reckase, 1974; Weiss, 1974; Patience, 1977), little or no literature has been written discussing operational characteristics of the procedure when program parameters and item pool attributes are varied. For this report, operational characteristics refer to how well the tailored testing procedure estimates a given true ability when provided an item pool and values for the controlling program parameters. The operational characteristics, item pool attributes, and program parameters will be described in detail shortly.

Although no literature was found which addressed the effects of varying program parameters, a few studies have appeared in the literature which investigated effects of item pool characteristics on the operation of tailored testing. Jensema (1975), for example, has investigated the influence of item pool size and item characteristics on a Bayesian tailored testing procedure. In general, Jensema found that when items are of adequate quality, it is not necessary to have very large item pools. Reckase (1976) concurred with Jensema in recommending a rectangular distribution of item pool difficulty values. In this latter study, the tailored testing procedure was based on an empirical maximum likelihood estimation of the ability parameter of the simple logistic (Rasch) model. Issues worthy of further investigation have surfaced in addition to item pool attributes, such as the effects of program parameters on the bias and variance of ability estimation. Bias and variance of ability estimation constitute the operational characteristics of the tailored testing procedure. Bias and variance of ability estimation and the program parameters will now be described more thoroughly.

Several articles have appeared in the literature which discuss bias of tailored testing ability estimation with regard to procedural bias toward subgroups of an examinee population such as minorities (see for example

---

Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, 1979. This research was supported by contract number N00014-77-C-0097 from the Personnel and Training Research Programs of the Office of Naval Research.

Pine and Weiss, 1978). The research reported here did not address this type of bias. Rather, ability estimate bias investigated for this paper concerned the failure of a maximum likelihood tailored testing procedure to result in an expected value of the ability estimate equal to a known true ability. In this sense, the attempt was to identify values for the program parameters, stepsize and acceptance range, as well as item pool characteristics which would provide the least statistical bias in ability estimation. The variance of ability estimates was the squared standard error of the ability estimates for a known true ability. The desire was to minimize this standard error. These two dependent measures provided the criteria for judging how well the tailored testing procedure estimated known abilities when the program parameters and item pool characteristics were varied.

PURPOSE

The primary purpose of the research described herein was to determine the operational characteristics of a one-parameter tailored testing procedure when program parameters and item pool attributes were varied. The program parameters investigated were the stepsize and acceptance range. The stepsize parameter specified the magnitude of movement of the ability estimate during the initial item selection phase of tailored testing. The acceptance range parameter determined how deviant the selected item's difficulty value could be from the requested item difficulty and still be acceptable for administration. Items were requested by the procedure to match the ability estimate which was computed after each item response. The item pool attributes varied were size, shape, and quality. Each of these variables will now be described more specifically.

Based on the premise of tailored testing that when an examinee answers an item correctly, the item administered should be more difficult and vice versa, the stepsize program parameter initially controlled how much more difficult or easy was the next item administered. The selection of items proceeded utilizing a fixed stepsize until the examinee had answered items both correctly and incorrectly. After both a correct and incorrect response had been obtained in the response string, a maximum likelihood ability estimate was obtained using an iterative search for the mode of the likelihood distribution. For a more complete description of the item selection and ability estimation components of this maximum likelihood tailored testing procedure see Patience, 1977. When an ability estimate had been obtained, items were selected from the pool to maximize the information function (Birnbaum, 1968). For the one-parameter model, the information function is maximized when the difficulty of the selected item equals the ability estimate. In the past, arbitrary values have generally been choosen for the stepsize. One of the primary goals of this research was to empirically investigate the effects of stepsize values on the bias and standard error of ability estimates. In so doing, the intent was to determine the optimal stepsize value which would minimize the bias and standard error of ability estimates.

The second program parameter investigated was the acceptance range. The acceptance range specified the amount of deviation in difficulty an administered item could have from the requested item difficulty and still be acceptable for administration. The acceptance range parameter monitored the appropriateness of items selected throughout the tailored test, i.e.,

both during item selection based on the fixed stepsize until both correct and incorrect responses had been obtained, and also during item selection to maximize the information function for a maximum likelihood ability estimate. If more than one item was within plus or minus the acceptance range of the desired item, the item with a difficulty value nearest the requested value was choosen. If no item was available from the pool within the specified acceptance range of the difficulty asked for, the tailored test was terminated. The primary aim then, regarding the acceptance range, was to determine what value or range of values yielded the least bias and standard error of ability estimates. Clearly, a small value for the acceptance range would have insured that items very near the desired item difficulty would be administered. On the other hand, too small an acceptance range value would have increased the chance of premature termination of the tailored test, which would have induced bias of the ability estimate. It should be noted that both stepsize and acceptance range interact with item pool attributes and, therefore, a choice of what values are optimal may not be made assuming independence of these controlling factors.

Item pool attributes studied in this research included size, shape, and quality. Item pools used in this investigation ranged in size from nine to one hundred and eighty-one items. Shapes of item pool distributions were normal, rectangular, bimodal, and skewed. Item pool quality referred to the contrast between actual and idealized pools. Idealized pools consisted of item difficulty parameters equally spaced from minus three to plus three with equal discrimination values of one and zero guessing as assumed by the Rasch model. Actual pools consisted of item difficulty values obtained from calibration runs using the Wright and Panchapakesan program based on the Rasch model (1969). In these pools, items were not equally spaced on the difficulty scale. It should be noted that, quite clearly, item pool attributes played a substantial role in the utility of the tailored testing procedure.

PROGRAMS

Two FORTRAN programs were used for investigating effects of program parameters and item pool attributes. The input variables for both programs included: a) acceptance range, b) stepsize, c) item pool size, d) item difficulty values for the various sizes and shapes of item pools, and e) the true abilities for which an estimate was to be made utilizing the program parameters and item pool provided. Both programs output the mean and standard deviation of the estimates of each true ability provided. These served as dependent measures for determination of the quality of estimation for the specific values of the acceptance range, stepsize, and item pool difficulties.

The first program, the TREE1P, was based on the concept of a propensity distribution. A propensity distribution in this context was defined as the probability distribution for observed ability estimates given a true ability, $P(\hat{\theta}|\theta)$ (Lord and Novick, 1968). The concept of a propensity distribution was extended from its use in true score theory to the context of latent trait ability estimation. The TREE1P program determined the propensity distribution for a given true theta, θ, analytically from the properties of the tailored testing model.

Briefly, the TREE1P program operated as follows. Initially an item of average difficulty was administered to the simulated examinee with known

true ability. Based on the probability function for the simple logistic model,

$$P\{\mu\} = \frac{e^{\mu(\theta - b)}}{1 + e^{(\theta - b)}} \qquad (1)$$

where μ is the item score (0 or 1), b is the item difficulty parameter, and θ is the ability parameter, the probability of a correct and the probability of an incorrect response was obtained. If the response was correct, the ability estimate was increased by the stepsize. If the response was incorrect, the ability estimate was decreased by the stepsize. Thus after one item was administered, two paths or branches were present on the "tree". (The tree diagram from probability theory was employed to represent the propensity distribution in this study.) Based on these first possible ability estimates, the closest items to each of the two estimates was selected for administration with the constraint that the difficulty of the items must have been within plus or minus the acceptance range from the present ability estimates. If no items were available to the program from the item pool provided, that branch was terminated at that point. However, assuming items were available, there existed four possible paths after the second item had been administered. As long as all correct or all incorrect responses were obtained on a given path, the ability estimates continued to be increased or decreased, respectively, by the stepsize. However, when both a correct and an incorrect response were present on a particular path of the tree, a maximum-likelihood ability estimation procedure obtained an ability estimate using an iterative search for the mode of the likelihood distribution.

Insert Figure 1 about here

To partially illustrate how the propensity distribution was determined by the TREE1P, Figure 1 shows a diagram representing the operation of the procedure on a nine item rectangular pool. The stepsize used for this illustration was 1.0 and the acceptance range was 0.3. The θ for this analytical derivation of the propensity distribution was set at zero. As was pointed out above, the procedure began by administering an item of average difficulty from the pool, i.e., the item with the difficulty parameter 0.0. The probability of a correct response, as determined by the probability function given above for the simple logistic model, was 0.5 and the probability of an incorrect response was 0.5.

After a correct response the ability estimate was increased by the stepsize, or after an incorrect response, it was decreased by the stepsize. Thus after one item, the ability estimate was either 1.0 with probability of 0.5 or -1.0 with a probability of 0.5. This procedure was followed so that finite ability estimates would be available after each item response, rather than the $\pm\infty$ value given by the maximum likelihood procedure. The expected value of the distribution after one item was 0.0 and the standard deviation was 1.0.

Based on these first possible ability estimates the closest items were selected from the pool with the restriction that they must have been within plus or minus 0.3 of the requested items. Thus, as Figure 1 illustrates, items with parameter estimates of plus and minus 0.75 were adminis-

tered to the estimated abilities plus and minus 1.00 respectively. On the upper branch of the tree, a correct response yielded an ability estimate that was again increased by the stepsize, since a maximum likelihood estimate could not be determined without both a correct and incorrect response. Now, the ability estimate was 2.0. The probability of a correct response to the item with the 0.75 difficulty parameter was the same except for the change in sign of the item parameters and ability estimates. When the item pool distribution being considered was symmetric, the results of the analyses were the same except for the change in sign.

Following the middle branches of the tree, an incorrect response to the item with difficulty 0.75 yielded an ability estimate of 0.375 from the maximum likelihood technique. The probability of this response was 0.68 based on the model. When the first item was missed and the second answered correctly, the probability of the second response was also 0.68. By the local independence assumption of the model, the probability of either a ± 2.0 estimate was $0.5 \times 0.32 = 0.16$ while the probability of ± 0.375 was $0.5 \times 0.68 = 0.34$. In this manner the propensity distribution after two items have been administered could be obtained. As noted at the bottom of Figure 1, the expected value was still 0.0 and the standard deviation (which was determined as the square root of the VAR(θ)) was 1.174.

The tree developed further in this same manner whenever items within the acceptance range were available. If all correct or incorrect responses were present, the fixed stepsize was used to make ability estimates. Once a mixture of correct and incorrect responses were present, the maximum likelihood ability estimate procedure was used. Note the "branches" of Figure 1 were "live" at ± 2.00 ability estimate but no items existed in the pool within ± 0.3 of the ability estimate ± 0.375. Therefore, those branches terminated.

The tree continues to develop by following all "live" paths. The program is finished after all branches are terminated by the condition that no items of appropriate difficulty are available in the pool. One may well imagine that as the number of items in the pool gets larger, the procedure is, practically speaking, bounded by the storage capacity of the computer facility and magnitude of one's computer budget. For the IBM 370/168 system on which the TREE1P program was run, it was found that sixty-one items was the practical upper limit on the number of items the pool could contain for any particular run of the various combinations of stepsize, acceptance range, and shape of the item difficulty distribution.

Due to the limitation on size of the item pool which could be investigated with the TREE1P program, the second computer program, SIM1P, was developed. This program was adapted from the tailored testing procedure based on the Rasch model which was already operational. This particular tailored testing procedure has been described thoroughly elsewhere (Reckase, 1974), so only the details pertinent to this research have been described here. The SIM1P program followed only one path for any given θ in contrast to the TREE1P. A particular path was selected using Monte Carlo simulation techniques. It provided for investigation of the properties of bias and variance of ability estimation with much larger item pools since the required storage and computation were substantially reduced as compared to the TREE1P program.

The following values served as input to the program: the stepsize, acceptance range, item pool difficulty values, θ, and number of simulated tests to be administered by the tailored testing procedure. The procedure

initially administered an item of average easiness from the pool of items provided. If a correct response was obtained, an item that was more difficult by a stepsize factor was administered. If an incorrect response was obtained, an easier item by the same value was administered. This fixed stepsize up and down procedure continued until both a correct and incorrect answer had been obtained in the response string. Then the procedure switched from the fixed stepsize procedure to maximum likelihood ability estimation. In both cases, items were selected to maximize the item information. Ability estimation was accomplished after each item was administered (provided correct and incorrect responses had previously occurred) by the maximum-likelihood estimation procedure using an iterative search for the mode of the likelihood distribution. The items administered had to be within plus or minus the acceptance range from the requested item difficulty. If no items were available within this range of the estimated ability, the procedure stopped. The only other stopping rule was based on a preset maximum number of items that was to be administered.

Items were scored correct or incorrect by the SIM1P program utilizing an internal random number generator. First, the probability of a correct response was computed using the formula for the probability function of the simple logistic model stated earlier. The theta for this computation was the true theta that was input into the program, and the difficulty parameter, b, was that of the item just administered to the simulated examinee. After this probability of a correct response had been determined, the random number generator selected a number between zero and one from a rectangular distribution. If this randomly selected number was less than or equal to the probability of a correct response, the item was scored correct. If the randomly selected number was greater than the probability of a correct response, the item was scored as incorrect. Provided both incorrect and correct responses had previously occurred, an ability estimate was made, and the next item administered was selected to maximize information for this estimated ability. This procedure continued until one of the stopping rules was encountered.

The major controlling program parameters for both the TREE1P and SIM1P were the stepsize and acceptance range values. The stepsize parameter controlled how quickly the procedure would move through the item pool while the acceptance range parameter specified how discrepant items could be from those desired and still be administered. The acceptance range also indirectly determined the number of items from the pool which were available for administration. Clearly, the wider was the acceptance range, the greater was the number of items that could have been chosen for administration.

The TREE1P and SIM1P programs used in this study for determining the optimal stepsize, acceptance range, item pool size, and item pool distribution were similar in that both output the mean and standard deviation of ability estimated for each true theta input. However, they differed in the manner in which the mean and standard deviation were determined. While the TREE1P pursued all possible paths through the item pool, the SIM1P followed only the path that was the result of the simulated interaction of an examinee with the tailored testing procedure. The mean and standard deviation from the TREE1P were actually expected values and square roots of variance computed from probabilities arising from the one-parameter

model and ability estimates arising from the maximum likelihood estimation technique. The SIM1P program provided a mean and standard deviation of the set of ability estimates obtained for each of the θs specified.

METHODS

To investigate the optimal stepsize, acceptance range, item pool size, and item pool shape, nearly all possible combinations of the following were input into the TREE1P and SIM1P programs for true abilities -3, -2, -1, 0, 1, 2, and 3. The stepsize values used were .3, .4, .5, .6, .693, .8, .9, 1.0, 1.5, 2.0, and 3.0, while acceptance ranges were .1, .2, .3, .4, and .5. Item pool sizes were 9, 13, 25, 31, 61, 72, 180, and 181. Item pool shapes investigated were normal, rectangular, bimodal, and skewed, with difficulty values constrained between plus and minus three. Idealized item pools (difficulty values in the above shapes with spacing dependent on shape and size of item pool) were constructed and used as input to the programs, as well as actual item pools (test items calibrated and formed into pools with less constraint on having items equally spaced along the difficulty scale).

The manner in which item pool size effects were investigated using simulations was to run the TREE1P and SIM1P programs on the various sized pools mentioned above. With the resulting data, plots and projections were made to estimate the item pool sizes needed for various accuracies of ability estimation.

The comparisons to determine the optimal combination of independent variables were based upon the mean and standard deviation of twenty-five simulated administrations of a tailored test to each θ using the SIM1P; where for the TREE1P program, the comparisons were of the expected value of θ, $E(\theta)$, and the standard deviation of θ, $\sqrt{Var(\theta)}$. Values of these dependent variables were compared across program runs using various sized item pools holding stepsize and acceptance range constant. They were also compared from runs using various shapes of item pools, holding size of item pool, stepsize, and acceptance range fixed. Additionally, comparisons were made of the dependent variables, first varying stepsize with all other variables fixed, and then varying the value of the acceptance range while holding all other variables constant. Since the TREE1P program was considered to yield the most accurate values, i.e. $E(\theta)$ and $\sqrt{Var(\theta)}$ based upon the propensity distribution, another comparison was deemed important. Because the SIM1P means and standard deviations were subject to sample variation, they were validated against values of the TREE1P for various runs on the sixty-one item pool. Also, the number of estimates of the true ability, i.e. the number of tailored tests administered to each simulated examinee by the SIM1P program was varied. This was done to check whether an appropriate number of administrations had been used.

RESULTS

The results of this study were to a great extent drawn from tables which summarized the results of the TREE1P and SIM1P programs. One issue to be investigated was to determine the distribution of item pool difficulty parameters that yielded the least bias and standard error of ability

estimates across the range of ability from -3 to +3. Another important question was how large an item pool was necessary to accomplish the goal of accurate ability estimation. Thirdly, a determination of the preferred magnitude of the stepsize parameter was desired. The fourth outcome of this study was to decide upon the approximate value of the acceptance range program parameter which would provide ability estimates with the least bias and standard error. These were the primary targets of the study.

Secondary goals of the study included a comparison of the performance of actual versus ideal item pools. The contrast of these has been previously described in the methods section. Another secondary objective was to compare the results of the TREE1P and SIM1P programs. In this regard, two concerns were investigated. One pertained to how close the SIM1P estimates of the means and standard deviations of ability were to the E(θ) and VAR(θ) determined by the TREE1P. The importance of this particular concern related to how well the SIM1P analyses on larger item pools provided reliable data on the primary questions of this study. It should be recalled that the motivation for development of the SIM1P program was to investigate the research questions of the study on larger item pools than the TREE1P program would realistically accommodate. The second concern subsumed under comparison of the TREE1P and SIM1P programs was to decide whether or not twenty-five estimates of each ability by the SIM1P was an adequate number. Several analyses were run of the SIM1P program on various item pools from which data had already been obtained from the TREE1P. By running the SIM1P on these pools and holding all other variables fixed except the number of test administrations, data were obtained pertaining to the adequacy of the SIM1P estimates of the means and standard deviations. Another matter along this same line was investigated with runs of the SIM1P on some of the larger pools. This was the question of whether or not twenty items as an upper limit of items administered by the tailored test was adequate.

Item Pool Shape

The TREE1P program (proportion distribution technique) was used to evaluate the effects of varying the shape of the item pool distribution on ability estimation. The rectangular item pools were obtained simply by selecting equally spaced items between +3.0 and -3.0 inclusive. The normal item pools were constructed such that the items were equally spaced in probability. That is, the area between item positions was kept constant in the range from +3.0 to -3.0 standard deviation units in the normal distribution. This procedure for producing the normally distributed pools has the effect of selecting more items around the difficulty value of zero and fewer items at the extremes. A similar procedure was used in selecting the item parameters for the bimodal pools as was used for selecting the normal pools. The negative half of the pool was centered around -.693 and the area under the normal distribution was used to place items around this point up to zero and down to -3.0. The same was true for the positive half of the pool. The reason $\pm$.693 were chosen as the two modes of the bimodal distribution was that, prior to the construction of a bimodal pool, .693 as a stepsize value had appeared promising. Therefore, after the first item was administered at 0, the stepsize of .693 would move the ability estimate out to one of the more dense regions of

the pool whether the examinee correctly or incorrectly answered the first item. The skewed item pool distribution of item parameters was constructed via a similar procedure to that for the normal and bimodal pools. That is, the items divided the distribution into equal areas. For the skewed pool, tables of the Pearson Type III distribution were used. The pool constructed was positively skewed (skewness = .5), and it should be noted that in any table included in the report, a skewed distribution always indicates a positive skew. However, the results would generalize to negatively skewed pools.

Results concerning the shape of the item pool distribution may be seen in Tables 1-6 for different combinations of values of the other variables. However, Tables 4 and 6 point out the more general trends of the item distribution study. In Table 6 the comparisons of the normal and rectangular pools of 25 items is shown for only acceptance ranges of 0.1 and 0.3 when paired with stepsizes of 0.5 and 0.7 respectively. These values of acceptance range and stepsize were chosen because they appeared to yield some of the least bias and least variance estimates. Specifically, the acceptance range of 0.1 was chosen to check whether the more dense item parameters near the middle of the normal distribution would make the use of the smaller acceptance range desirable.

As can be seen from Table 6, the normal distribution appears to be inferior to the rectangular item distribution in almost all cases. Except for the 0.1 acceptance range data at 0.5 and 1.0 ability levels, either the expected values deviate more from θ or the standard deviations are larger, or both. It is interesting to note that even the estimates at ability 0.0 are not as good for the normally distributed pool as for the rectangular pool, even though more items are present for estimation of ability.

Table 4 shows the expected values and standard deviations from the TREE1P on normal, bimodal, rectangular, and positively skewed pools. Each of these contained sixty-one items. The stepsize was fixed at 0.693, and the acceptance range was held at 0.30 for all runs. Again the rectangular pool performed better overall than did the other shapes of item difficulty distributions. For true abilities zero and one, the standard deviation of ability estimates, as well as the bias of the estimates, was smallest for analyses using the rectangular pool. At the ability levels of two and three, the rectangular pool yielded estimates with less bias in the expected values but larger standard deviations than the other shaped pools.

The results obtained from the TREE1P would have been the same for the negative end of the ability continuum when the pools were symmetric. Therefore, only the positive values of ability were run for the normal, rectangular and bimodal pools. However, for the skewed pool containing sixty-one items, the negative ability values of -1, -2, and -3 were run using the same program parameters as were indicated in Table 4. The results were as follows. For -1, the $E(\theta) = -1.189$ and $S_\theta = 0.836$. For -2, the $E(\theta) = -2.249$ and $S_\theta = 0.761$. For -3, the $E(\theta) = -2.935$ and $S_\theta = 0.577$. Even if one considered this skewed pool as being better suited for ability levels around minus two to minus one, since it contained more items around that region, it did not perform better than the rectangular pool.

Overall conclusions about the most preferred item distribution were that the rectangular pool was most apt to yield the least bias and smallest standard error of ability estimates across the ability scale. One important

caution was suggested by some of the results. When setting up an item pool for use with tailored testing procedures (especially those having a comparable parameter to the acceptance range), it is important to look carefully at the frequency distribution to be assured that no substantial gaps exist in any area of the continuum. Otherwise, one may expect poor estimation of ability at that region on the continuum. In this regard, one should view the estimates of true ability 3.0 as understandably limited, in as much as the item pools did not have any items beyond difficulty 3.0. For best estimation of ability, the pool should have a dense uniform distribution of items around the ability level to be estimated.

Item Pool Size

The criteria for judging how large an item pool was needed for the tailored testing procedure were again the bias and standard error of ability estimates. The results of the simulations using both the TREE1P and SIM1P programs have been condensed, and the general trend has been illustrated in Figure 2. The values of the $E(\theta)$ and S_θ which have been plotted for item pools of size 9, 13, 25, 31, and 61 were obtained from the TREE1P. Each of these pools had a rectangular distribution of item difficulty parameters. The item pools with 72 and 180 items were actual tailored testing item pools. The pool of 72 items consisted of item difficulty parameters from the calibration of a set of vocabulary items. This pool was named VC1PL. The pool with 180 items was constructed using item difficulty parameters resultant of calibrations of items covering the evaluation techniques portion of an introductory measurement and evaluation course. This pool was known as ET1PL. The distributions of item difficulty for VC1PL and ET1PL were graphed in Figure 3 and Figure 4 respectively. The means and standard deviations of ability estimates on the SIM1P runs on these latter pools have been included in the plots of Figure 2. Each analysis represented in this figure had θ set equal to 1.0, the stepsize fixed at 0.693, and acceptance range equal to 0.30.

The top graph of Figure 2 illustrates that as item pool size reaches 61 for this particular set of analyses, the $E(\theta)$ is equal to θ. The bias of the ability estimates is essentially zero. The bottom graph of Figure 2 shows that as item pool size increases, the standard error becomes less. While these plots should be considered as rough approximations of the relationship between item pool size and ability estimate bias and standard error, the indication appears to be that with a uniform distribution of item difficulty, $\theta = 1$, and the program parameters equal to the values used here, one could expect very little bias and a standard error of about 0.1 with an item pool consisting of around 200 items. More will be presented on item pool size in the discussion section of this paper.

Stepsize

The results of the study of the preferred magnitude of the stepsize program parameter may be seen in Tables 1, 2, 3 and 7. Tables 1, 2, and 3 give the $E(\theta)$ and S_θ from TREE1P analyses of $\theta = 0, 1, 2$, and 3 using item pools of size 9, 13, 25, 31, and 61. Tables 1, 2, and 3 show results for the rectangular, normal, and bimodal distributions of item difficulty parameters respectively. Table 7 presents the results of the SIM1P analyses on the ET1PL item pool for $\theta = -3, -2, -1, 0, 1, 2$, and 3. Negative θ

values are not shown in Tables 1, 2, and 3 since the results of the TREE1P on the pools used are the same as for the positive θ values except for the change of sign on the $E(\theta)$s. This should be expected since the item pool distributions of item difficulty are symmetric around zero. The acceptance range for all analyses for Tables 1, 2, and 3 was 0.30. For the SIM1P analyses of the ET1PL, a substantially larger item pool, a smaller acceptance range, 0.25, was used as is noted at the bottom of Table 7. Another variable recorded in Table 7 is the mean number of items administered for the 25 tests simulated by the SIM1P for each ability level. The maximum number of items per simulated test was 20 for these SIM1P analyses.

In general, results presented in Tables 1, 2, and 3 suggest that stepsizes between 0.5 and 1.0 give fairly unbiased estimates, and also have the smallest standard errors. Larger stepsizes tend to have a positive bias and larger standard errors. From several graphs like the ones presented in Figure 5, the stepsize value of 0.693 appears to be the best overall compromise value which achieves less bias while holding the standard error down. Figure 5 shows the $E(\theta)$ and S_θ for the 31 item rectangular, normal, and bimodal pools when $\theta = 1.0$ and the acceptance range equals 0.30. Due to the cost of running the TREE1P on larger item pools, not all cells of Tables 1, 2, and 3 for the 61 item pools have been analyzed.

Table 7 of the SIM1P on the ET1PL pool suggests that a stepsize between 0.4 and 0.7 is probably better for less bias and standard error. It should be recalled that the SIM1P is subject to sample variation, but in general, the results seem to suggest that a stepsize of about 0.7 is appropriate. However, a trend which should be investigated further is that larger item pools seem to do better with smaller stepsizes and vice versa.

Acceptance Range

The results of the acceptance range study are given in Tables 8, 9, and 10. Table 8 presents the $E(\theta)$ and S_θ for stepsizes 0.5, 0.693, 1.0, and 1.5; acceptance ranges 0.1, 0.2, 0.3, and 0.4; and ability levels 0.0, 1.0, 2.0, and 3.0 from TREE1P analyses. All of the results in Table 8 are based on the 25 item rectangular pool. From Table 8 it can be seen that in most cases, as the acceptance range increases, the standard deviation decreases. This is a reasonable result since more items are available for administration with a larger acceptance range. However, there is also a trend present in the amount of bias in estimate as the acceptance range increases, particularly at the higher ability levels and for the larger stepsizes.

Table 9 shows the results of the SIM1P on the VC1PL pool; 25 test administrations per ability level; 20 item upper limit; stepsize = .693; and $\theta = -3, -2, -1, 0, 1, 2,$ and 3. The mean number of items is also indicated. These results indicate that an acceptance range of 0.30 is probably the best compromise value for minimizing bias and standard error of ability estimates across the range of θ. Table 10 shows the results of the SIM1P on the ET1PL pool; 25 test administrations per ability level; 40 item upper limit; stepsize = .693; and $\theta = -3, -2, -1, 0, 1, 2,$ and 3. Again, the mean number of items is indicated. These results on ET1PL are somewhat more ambiguous although the extreme acceptance range values are clearly inferior to the more moderate values of .2 to .4. In cases such as this, one should consider a combination of the density of the

item pool across the range of θ and whether a particular θ range should be estimated more precisely than others, in order to decide on the best acceptance range value. Decisions regarding the best value of program parameters cannot be made independent of considerations such as the size and shape of the item pool to be used.

Secondary Results

Secondary results include the comparison of the performance of actual versus ideal item pools previously discussed. Table 11 shows this comparison, and overall, the ideal pool did not perform much better than the ET1PL pool.

Another comparison was of the SIM1P and TREE1P programs on the same pools using the same program parameter values. By looking at Table 4 and Table 5, one may see that the SIM1P did a reasonably good job of approximating the TREE1P results at θ = 2 for the bimodal and skewed pools. Also, from Table 5, it can be seen that increasing the number of tests administered by the SIM1P did not dramatically change the means and standard deviations. Therefore, 25 administrations seemed adequate.

Finally, by comparing cells of Tables 7 and 10, one can see that increasing the maximum number of items administered from 20 to 40 does not substantially change the means and standard deviations from the SIM1P. This comparison is not exact because the acceptance range of 0.25 used for analyses in Table 7 does not precisely equal the value of 0.2 or 0.3 for acceptance range in Table 10. Neither is the stepsize of 0.7 in Table 7 exactly equal to 0.693 used in Table 10. However, the values seemed close enough to make a comparison, and the result of this comparison seemed to indicate that 20 items as an upper limit was adequate. Note that the mean number of items recorded in both tables illustrated that the procedure approached the upper limit in the middle range of θ.

DISCUSSION

TREE1P

A possible explanation for the larger standard deviation given by analyses run on the rectangular pool at the more extreme values of the ability continuum was suggested by a close look at the development of the propensity distribution by the TREE1P for the various shaped item pools.

A property of the TREE1P and the manner in which it developed the propensity distributions was that the standard deviation actually increased as more branches or levels resulted from items administered to more and more possible ability estimates. This increase of the standard deviation of ability estimates stabilized and converged for the smaller item pools as the paths or branches of the "tree" terminated. For the larger pools (especially the sixty-one item pools), the standard deviation initially increased but as branches were terminated the standard deviation came down. This pattern of increasing standard deviation of ability estimates during the early formulation of the propensity distribution was evident for all shapes of the distributions of items in the pools.

However, the patterns of convergence to the final standard deviations yielded by the TREE1P were different for the various shapes of item pools

at different ability levels. Tables 1, 2, and 3 show a general tendency for the standard deviations of ability estimates of the true abilities zero and one to be larger for the normal and bimodal pools than for the rectangular pools. But for ability levels two and three, the standard deviations of ability estimates were generally larger for the rectangular pools than for the normal and bimodal pools. This trend was consistent across most of the TREE1P analyses. The explanation proposed was that, because more items were available for administration to the more extreme levels of ability (i.e. 2 and 3) when the rectangular pool was used, the standard deviation of ability estimates was larger since the standard error was more accurately estimated. The standard deviations of the estimates from the normal and bimodal pools for these true ability levels were smaller, since paths or branches were often terminated because no items were available within the acceptance range of the estimated abilities. In short, when fewer items were in the pool around a particular true ability, there were fewer paths allowed to develop in the propensity distribution due to the stopping rules. Therefore, the standard deviation of ability estimates at that particular level was an underestimate. A logical check for this phenomenon was the prediction that when the acceptance range was made smaller, the drop in standard deviations for the more extreme ability levels would be more pronounced with the normal pool than for the rectangular. This did appear to be the case. The point is that the smaller standard deviations for ability levels two and three yielded by the TREE1P when normal or bimodal pools were used probably should not be weighted too heavily, as the tendency appears to be somewhat of an artifact of the procedure. The values obtained for the rectangular pools may well be more representative.

SIM1P

SIM1P was designed to score and administer items in the manner previously described based on the rationale that this approach was a reasonable simulation of the behavior of an examinee when interacting with a tailored test. The pseudo examinee with some specified true ability was presented an item of average difficulty from the pool, because, given we have no prior information about his ability, the best guess of an item appropriate for the examinee was one of average difficulty. Scoring of each item using the examinee's θ in the one-parameter formula and then selecting a random number from a rectangular distribution between zero and one was deemed a reasonable simulation assuming the one-parameter model was correct. Clearly, the larger the probability of a correct response was, the greater the chance was that the random number generated was less than or equal to the probability specified by the model of a correct response. However, there was ample provision for the reality that occasionally an examinee with adequate ability to answer an item correctly will still respond incorrectly and vice versa. While the probability of a correct response was computed using the examinee's true theta, item selection procedures used the fixed stepsize until correct and incorrect responses were present, and then selected items maximizing information for the estimated ability. This approach rounded out the simulation of the interaction between examinee and tailored test with respect to the SIM1P.

Item Pool Size

The methods employed for the investigation of the effects of item pool size on the operation of the one-parameter maximum likelihood tailored testing procedure were simulations, but theoretical methods have also been proposed. Lord (1970) suggested a formula for the number of items required for a fixed stepsize procedure (selecting items more difficult by the stepsize when correct responses were given and vice versa). The formula is

$$N = (1 + R/d)\ (n - R/2d) \qquad (2)$$

where $\pm R$ is the range of item difficulties desired, d is the stepsize and a submultiple of R, and n is the maximum number of items to be administered. For example, if R were plus three to minus three, d were set at 0.5, and n were twenty, the formula would give

$$119 = (1 + 3.0/0.5)\ (20 - 3.0/(2 \times 0.5)). \qquad (3)$$

With this set of values, 119 items would be required if the exact item requested were to be available.

This formula does not directly apply to some tailored testing procedures which use a variable rather than a fixed stepsize. Also, most testing procedures allow administration of slightly discrepant items from those requested by the procedure (the acceptance range specifies how discrepant). Procedures using a variable stepsize tend to require more items because, as the procedure converges to an ability estimate, the stepsize in effect becomes smaller and smaller. Allowing items to be administered which differ slightly from the requested item compensates to an extent for the increase in number of items caused by the variable stepsize. Another limitation of the formula is that several tailored testing procedures administer items until a specified precision is reached instead of using a preset maximum number of items as a stopping rule.

Another theoretical method of estimating how large an item pool should be is to determine the number of items required to reach a specified precision of ability estimation, given that equally spaced, perfectly discriminating items are available. With these ideal or optimal circumstances, the precision of an ability estimate is equal to the difference between adjacent items. For example, an item pool with seven equally spaced items from -3.0 to +3.0 would classify examinees into categories 1.0 scale unit apart. The number of item responses required to make the classification would be

$$k = \log_2 n \qquad (3)$$

where n is the size of the item pool, since 2^k is the number of branches in the tree diagram after k items are administered. By specifying the precision desired, e, the minimum item pool size can be determined by the range of ability, R, divided by e, plus one.

$$n = \frac{R}{e} + 1 \qquad (4)$$

The minimum number of items administered to classify all ability levels in the tailored testing situation is

$$k = \log_2[\frac{R}{e} + 1]. \quad (5)$$

Some results obtained by the application of the formulas based on the theoretical method for estimating the number of items needed in a pool, given the precision desired, have been indicated in Table 12. The requirements for pool size were computed for the range of ability, -3.5 to +3.5, given the desired classification interval size. As has been pointed out, these results are for a rectangular pool of hypothetical items with perfect discrimination and zero guessing probabilities. With these restrictions, the item pool sizes shown must be regarded as lower limits. The minimum session length indicates the fewest number of items that would have to be administered in order to classify an ability level within the capabilities of the item pool. These also are based on hypothetically perfect items and item pools, and should be considered as lower limits. The values in the column labelled simulated length are the number of items required to reach a best estimate using the most likely response pattern simulated. All results in this column are based on $\theta = 0.0$.

In some cases the simulated session length is less than the minimum predicted length because of the choice of ability level. Setting the stepsize equal to 0.693 tends to keep the process near the middle of the item pool, speeding up convergence for abilities near 0.0. If an ability of 3.0 had been used, the session length would have been 6, well over the minimum predicted values. Thus, the minimum session length refers to the number of items needed across the ability range, and under specified circumstances fewer items may be required.

These results using simulated tests have been compared to actual tailored testing convergence plots and found to be fairly good approximations (Reckase, 1976). One observation of importance is that, from convergence plots, it can be seen that giving too many easy items causes bias in ability estimation. Reckase (1975) has discussed this effect in detail.

Generalizability

It should be kept in mind that this report focused primarily on program parameters and item pool attributes as they interacted with the one-parameter maximum likelihood tailored testing procedure currently in operation for this research project. Clearly, the inferences drawn from the results should generalize to other tailored testing applications using similar conceptual formulations of operation. In this sense, the results of this study were intended not as isolated studies of item pool size and shape, stepsize magnitude, and value of the acceptance range, but rather intended to generalize to fairly concrete statements about the preferred operation of a one-parameter tailored testing procedure. As was expected, item pool attributes and program parameters interacted to a great extent in the determination of the degree of bias and amount of variance in ability estimation. The intention in drawing up the numerous tables and figures of this report was to illuminate trends of interaction among these variables. These trends,

in large part, were the primary thrust of this report. They should be helpful in applying tailored testing procedures in which some of the variables, such as item pool attributes, have been fixed by practicality.

In conclusion, this paper was intended as a guide for those setting up a tailored testing procedure. The paper does not, by any means, exhaust all the inferences that could be drawn from this set of data. The report presented here has been an attempt to condense a more elaborate technical report which is presently being developed. One point should be made in closing. This strategy for investigating bias and standard error was motivated by the need to determine these values across the ability continuum, since our efforts were directed toward developing a criterion-referenced tailored test. In criterion referenced testing, it is essential to know the effects of estimate bias and standard error on decisions made at various points along the ability continuum.

References

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Jensema, C. J. Bayesian tailored testing and the influence of item bank characteristics. Paper presented at the Invitational Conference on Adaptive Testing, Washington, D.C., 1975.

Lord, F. M. Some test theory for tailored testing. In W. H. Holtzman (Ed.), Computer-assisted instruction, testing, and guidance, New York: Harper and Row, 1970.

Lord, F. M. and Novick, M. R. Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Patience, W. M. Description of components in tailored testing. Behavior Research Methods and Instrumentation, 1977, 9(2), 153-157.

Pine, S. M., and Weiss, D. J. A comparison of the fairness of adaptive and conventional testing strategies (Research Report 78-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1978.

Reckase, M. D. An interactive computer program for tailored testing based on the one-parameter logistic model. Behavior Research Methods and Instrumentation, 1974, 6(2), 208-212.

Reckase, M. D. The effect of item choice on ability estimation when using simple logistic tailored testing model. ERIC Document Reproduction Service No. ED 106 342, 1975.

Reckase, M. D. The effect of item pool characteristics on the operation of a tailored testing procedure. Paper presented at the spring Meeting of the Psychometric Society, Murray Hill, New Jersey, 1976.

Weiss, D. J. Strategies of adaptive ability measurement. Research Report 74-5, Department of Psychology, University of Minnesota, December, 1974.

Wright, B. D. and Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.

Table 1
Expected Values and Standard Deviations
from TREE1P on Rectangular Item Pools
Varying Pool Size, Step Size and Ability Level

Pool Size	Step Size	Ability Level 0		1		2		3	
		E(θ)	S_θ	E(θ)	S_θ	E(θ)	S_θ	E(θ)	S_θ
	0.5	-0.000	0.645	0.405	0.603	0.709	0.482	0.877	0.335
9	0.693	-0.001	1.025	0.756	1.113	1.593	1.217	2.388	1.139
	1.0	-0.001	1.155	0.821	1.213	1.685	1.298	2.548	1.286
	1.5	-0.001	1.192	0.934	1.268	1.966	1.439	3.016	1.423
	0.5	-0.001	0.765	0.655	0.937	1.577	1.219	2.599	1.201
13	0.693	-0.001	0.976	0.733	1.056	1.587	1.217	2.454	1.168
	1.0	-0.001	1.187	1.037	1.150	1.995	1.085	2.822	1.005
	1.5	-0.006	1.125	0.899	1.249	1.960	1.463	3.045	1.424
	0.25	-0.001	0.547	0.584	0.809	1.606	1.200	2.783	1.190
	0.5	-0.001	0.736	0.857	0.842	1.933	1.000	2.964	0.809
	0.6	0.001	0.744	0.896	0.888	1.986	1.004	2.955	0.788
	0.693	-0.013	0.786	0.910	0.892	1.984	0.980	2.925	0.765
25	0.8	-0.013	0.801	0.931	0.934	2.047	1.042	3.045	0.845
	0.9	-0.001	0.845	0.996	0.895	2.061	0.972	2.996	0.784
	1.0	-0.001	0.829	0.990	0.901	2.099	1.036	3.135	0.867
	1.5	-0.001	0.972	1.109	1.086	2.318	1.221	3.389	1.040
	1.7	-0.001	1.473	1.329	1.417	2.477	1.116	3.143	0.614
	2.0	-0.001	1.551	1.389	1.553	2.673	1.348	3.535	0.846
	3.0	-0.001	1.555	1.361	1.741	2.863	1.930	4.248	1.750
	0.5	0.004	0.726	0.949	0.788	2.022	0.902	3.018	0.725
31	0.693	-0.003	0.742	0.973	0.826	2.068	0.907	2.997	0.672
	1.0	-0.003	0.776	1.009	0.866	2.140	0.995	3.183	0.817
	1.5	-0.005	0.925	1.116	1.050	2.002	1.388	3.382	1.023
	0.5	-0.001	0.598	0.989	0.657	2.116	0.804	3.133	0.593
61	0.693	-0.001	0.610	1.008	0.677	2.138	0.805	3.111	0.566
	1.0	-0.000	0.641	1.039	0.745	2.229	0.915	3.289	0.689
	1.5								

Note. Acceptance Range = 0.30

Table 2
Expected Values and Standard Deviations
from TREElP on Normal Item Pools
Varying Pool Size, Step Size and Ability Level

Pool Size	Step Size	Ability Level 0		1		2		3	
		E(θ)	S_θ	E(θ)	S_θ	E(θ)	S_θ	E(θ)	S_θ
	0.5	-0.001	1.018	0.848	0.847	1.318	0.491	1.463	0.226
9	0.693	-0.001	1.098	0.960	0.966	1.601	0.655	1.898	0.382
	1.0	-0.001	1.269	0.880	1.084	1.641	0.632	1.877	0.334
	1.5	0.000	1.500	0.693	1.330	1.142	0.972	1.358	0.638
	0.5	-0.001	1.028	1.062	0.866	1.697	0.514	1.922	0.237
13	0.693	-0.001	1.101	1.002	0.942	1.648	0.628	1.932	0.358
	1.0	-0.000	1.273	1.146	1.020	1.760	0.548	1.946	0.231
	1.5	-0.001	1.439	1.272	1.282	2.219	1.188	3.031	1.258
	0.25	-0.001	0.847	1.110	0.858	1.969	0.576	2.210	0.408
	0.5	-0.001	0.891	1.184	0.837	2.016	0.572	2.359	0.278
	0.6	-0.001	0.980	1.203	0.847	1.965	0.528	2.263	0.266
	0.693	-0.000	0.956	1.174	0.811	1.871	0.482	2.079	0.227
25	0.8	-0.001	1.009	1.234	0.871	2.004	0.539	2.292	0.253
	0.9	-0.001	1.052	1.290	0.964	2.223	0.784	2.818	0.658
	1.0	-0.001	1.055	1.295	0.979	2.263	0.858	2.949	0.820
	1.5	-0.001	1.327	1.384	1.186	2.394	1.070	3.167	1.114
	1.7	-0.001	1.536	1.521	1.363	2.549	0.968	3.047	0.628
	2.0	-0.001	1.738	1.653	1.600	2.845	1.248	3.492	0.884
	3.0	-0.001	1.792	1.627	1.749	2.928	1.814	4.045	1.883
	0.5	-0.000	0.869	1.218	0.805	2.046	0.557	2.385	0.277
31	0.693	-0.001	0.964	1.268	0.880	2.192	0.734	2.778	0.607
	1.0	-0.001	1.018	1.323	0.951	2.300	0.823	2.969	0.787
	1.5	-0.001	1.301	1.404	1.155	2.410	1.043	3.176	1.092
	0.5								
61	0.693	-0.000	0.866	1.256	0.873	2.267	0.693	2.840	0.543
	1.0								
	1.5								

Note. Acceptance Range = 0.30

Table 3
Expected Values and Standard Deviations
from TREE1P on Bimodal Item Pools
Varying Pool size, Step Size and Ability Level

Pool Size	Step Size	Ability Level 0		1		2		3	
		E(θ)	S_θ	E(θ)	S_θ	E(θ)	S_θ	E(θ)	S_θ
	0.5	-0.004	1.020	0.231	0.443	1.312	0.495	1.473	0.245
9	0.693	-0.004	1.095	0.951	0.968	1.601	0.666	1.903	0.383
	1.0	-0.001	1.264	1.036	1.042	1.639	0.628	1.876	0.331
	1.5	-0.001	1.442	1.216	1.326	2.187	1.252	3.027	1.291
	0.5	-0.001	1.006	1.009	0.903	1.671	0.579	1.917	0.275
13	0.693	-0.001	1.104	1.001	0.945	1.647	0.630	1.932	0.358
	1.0	-0.000	1.267	1.143	1.011	1.754	0.551	1.946	0.238
	1.5	-0.000	1.436	1.274	1.276	2.217	1.181	3.029	1.252
	0.25	-0.000	0.920	1.102	0.855	2.001	0.623	2.264	0.421
	0.5	-0.001	0.870	1.152	0.867	2.024	0.594	2.373	0.278
	0.6	-0.001	0.951	1.173	0.875	1.976	0.536	2.271	0.242
	0.693	-0.001	0.964	1.207	0.933	2.174	0.768	2.774	0.612
25	0.8	-0.001	0.953	1.183	0.887	2.020	0.589	2.335	0.272
	0.9	-0.002	1.025	1.260	0.994	2.246	0.780	2.833	0.631
	1.0	-0.001	1.017	1.257	1.002	2.280	0.860	2.969	0.791
	1.5	-0.001	1.294	1.350	1.192	2.396	1.064	3.176	1.091
	1.7	0.002	1.491	1.483	1.362	2.543	0.959	3.047	0.612
	2.0	-0.000	1.717	1.609	1.592	2.831	1.235	3.485	0.871
	3.0	-0.001	1.761	1.601	1.763	2.953	1.803	4.070	1.857
	0.5	-0.000	0.796	1.145	0.816	2.060	0.621	2.476	0.406
31	0.693	-0.000	0.924	1.229	0.912	2.218	0.741	2.814	0.585
	1.0	-0.000	0.957	1.262	0.956	2.298	0.832	3.004	0.758
	1.5	-0.002	0.968	1.284	1.049	2.446	1.080	3.338	1.015
	0.5	0.006	0.726						
61	0.693	-0.000	0.857	1.245	0.876	2.281	0.688	2.852	0.525
	1.0	0.033	0.867						
	1.5	0.185	1.128						

Note. Acceptance Range = 0.30

Table 4
Expected Values and Standard Deviations from TREE1P on Various Shaped Item Pools

Pool Shape	Ability Level							
	0		1		2		3	
	E(θ)	S_θ	E(θ)	S_θ	E(θ)	S_θ	E(θ)	S_θ
normal	-0.000	0.866	1.256	0.873	2.267	0.693	2.840	0.543
bimodal	-0.000	0.857	1.245	0.876	2.281	0.688	2.852	0.525
rectangular	-0.001	0.610	1.008	0.677	2.138	0.805	3.111	0.566
skewed	0.040	0.815	1.282	0.858	2.257	0.670	2.801	0.561

Note. All runs were on pools with 61 items with the stepsize and acceptance range program parameters set at 0.693 and 0.30 respectively.

Table 5
Means and Standard Deviations from SIM1P on a Bimodal and Skewed Item Pool Varying Number of Test Administrations

Number of Tests Administered	Shape of Pool			
	Bimodal		Skewed	
	$\bar{X}_\theta$	S_θ	$\bar{X}_\theta$	S_θ
25	2.207	0.627	2.193	0.622
50	2.242	0.634	2.225	0.627
75	2.262	0.645	2.216	0.603

Note. All runs made with 20 item upper limit, stepsize = .693, and acceptance range = 0.30. The true ability was set at 2.0. Both the pools had 61 items.

Table 6
Comparison of TREE1P Results from
25 Item Rectangular and Normal Item Distributions

Acceptance Range	Step Size	Distribution Shape	Ability Level 0.0		0.5		1.0		2.0		3.0	
			$E(\theta)$	S_θ	$E(\theta)$	S_θ	$E(\theta)$	S_θ	$E(\theta)$	S_θ	$E(\theta)$	S_θ
0.1	0.5	R	-0.001	0.918	0.470	0.927	0.944	0.943	1.893	0.968	2.764	0.884
		N	-0.009	0.951	0.522	0.904	0.980	0.762	1.468	0.426	1.555	0.251
0.3	0.7	R	-0.013	0.787	0.430	0.824	0.911	0.893	1.986	0.984	2.933	0.773
		N	-0.000	0.959	0.623	0.922	1.169	0.821	1.877	0.491	2.093	0.231

Table 7
Means and Standard Deviations
from SIM1P on ET1PL Item Pool
Varying Stepsize

Stepsize		Ability Level						
		-3	-2	-1	0	1	2	3
.1	$\bar{X}_\theta$	-2.886	-2.145	-0.992	-0.050	1.135	1.991	3.331
	S_θ	0.715	0.728	0.486	0.534	0.502	0.627	0.788
	Mni*	13.04	15.88	19.24	20.00	20.00	19.84	18.40
.2	$\bar{X}_\theta$	-2.779	-2.230	-1.132	0.129	0.952	2.009	2.972
	S_θ	0.491	0.681	0.550	0.461	0.374	0.515	0.857
	Mni*	12.24	13.96	19.68	20.00	20.00	19.76	18.24
.3	$\bar{X}_\theta$	-3.157	-2.139	-1.134	0.064	1.018	2.055	3.213
	S_θ	0.652	0.645	0.800	0.503	0.363	0.516	0.844
	Mni*	10.04	14.48	18.56	20.00	19.92	19.56	16.08
.4	$\bar{X}_\theta$	-3.168	-2.250	-1.052	0.001	1.070	1.987	2.910
	S_θ	0.611	0.782	0.547	0.518	0.444	0.531	0.554
	Mni*	9.56	17.04	19.24	20.00	20.00	19.48	18.12
.5	$\bar{X}_\theta$	-2.762	-2.096	-1.122	-0.070	1.136	2.076	3.053
	S_θ	0.539	0.619	0.700	0.539	0.562	0.548	0.718
	Mni*	9.20	14.72	18.12	20.00	20.00	19.40	16.28

Note: All runs made with 25 administrations per ability level, 20 item upper limit, and acceptance = .25.

*Mni = mean number of items administered.

Table 7 (Cont.)
Means and Standard Deviations
from SIM1P on ET1PL Item Pool
Varying Stepsize

Stepsize		Ability Level -3	-2	-1	0	1	2	3
.7	$\bar{X}_\theta$	-3.061	-2.175	-1.026	-0.065	1.029	1.950	2.913
	S_θ	0.561	0.460	0.573	0.469	0.516	0.696	0.533
	Mni*	7.80	13.12	18.92	20.00	20.00	19.16	15.92
.9	$\bar{X}_\theta$	-3.134	-2.271	-1.241	0.094	0.959	2.029	3.310
	S_θ	0.499	0.790	0.898	0.419	0.380	0.531	0.799
	Mni*	5.92	11.40	16.96	20.00	19.84	19.20	13.28
1.5	$\bar{X}_\theta$	-3.739	-2.501	-1.389	0.101	1.035	2.437	3.239
	S_θ	0.876	0.961	0.910	0.598	0.792	1.118	1.010
	Mni*	5.32	10.80	18.04	20.00	19.32	16.16	12.72
2.0	$\bar{X}_\theta$	-3.683	-2.972	-1.482	-0.329	1.100	2.032	3.631
	S_θ	0.514	1.044	1.194	1.175	0.450	0.913	1.345
	Mni*	4.24	8.76	16.56	18.56	19.96	18.48	13.36
3.0	$\bar{X}_\theta$	-4.530	-2.942	-1.751	-0.042	1.230	2.511	4.471
	S_θ	1.591	1.494	1.916	0.465	1.117	1.556	1.519
	Mni*	5.04	10.68	16.52	20.00	19.28	17.04	8.60

Note. All runs made with 25 administrations per ability level, 20 item upper limit, and acceptance range = .25.

*Mni = mean number of items administered

Table 8

Expected Values and Standard Deviations
from TREE1P on 25 Item Rectangular Pool
by Step Size and Acceptance Range

Ability Level	Acceptance Range	Step Size 0.5		0.693		1.0		1.5	
		$E(\theta)$	S_θ	$E(\theta)$	S_θ	$E(\theta)$	S_θ	$E(\theta)$	S_θ
0.0	.1	-0.00	0.92	-0.00	0.84	-0.00	1.04	-0.01	1.07
	.2	-0.00	0.81	-0.02	1.01	-0.00	0.88	-0.00	1.06
	.3	-0.00	0.74	-0.01	0.79	-0.00	0.83	-0.00	0.97
	.4	-0.00	0.76	-0.01	0.78	-0.00	0.81	-0.00	0.93
1.0	.1	0.94	0.94	0.55	0.80	1.08	1.06	0.89	1.23
	.2	0.89	0.87	1.00	1.04	1.00	0.94	0.90	1.22
	.3	0.86	0.84	0.91	0.89	0.99	0.90	1.11	1.09
	.4	0.94	0.81	0.96	0.83	1.00	0.89	1.10	1.07
2.0	.1	1.89	0.97	0.97	0.66	2.09	1.03	1.99	1.45
	.2	1.92	0.99	1.88	0.97	2.08	1.04	2.00	1.45
	.3	1.93	1.00	1.98	0.98	2.10	1.04	2.32	1.22
	.4	2.01	0.92	2.03	0.91	2.12	1.02	2.33	1.21
3.0	.1	2.76	0.88	1.21	0.46	2.93	0.93	3.09	1.39
	.2	2.89	0.85	2.74	0.89	3.08	0.91	3.10	1.39
	.3	2.96	0.81	2.92	0.76	3.14	0.87	3.39	1.04
	.4	3.00	0.74	2.97	0.72	3.16	0.84	3.42	1.01

Table 9
Means and Standard Deviations
from SIM1P on VC1PL Item Pool
Varying Acceptance Range

Acceptance Range		Ability Level						
		-3	-2	-1	0	1	2	3
.1	$\bar{x}_\theta$	-1.938	-1.713	-0.994	-0.491	1.101	1.873	2.913
	s_θ	0.430	0.794	0.573	0.810	0.976	0.676	0.447
	Mni*	3.64	5.56	6.84	8.52	11.12	9.72	6.24
.2	$\bar{x}_\theta$	-2.747	-2.133	-1.193	-0.152	1.208	2.268	2.889
	s_θ	0.790	0.520	0.779	0.544	0.739	0.686	0.540
	Mni*	6.96	8.88	12.56	14.44	15.44	10.56	7.20
.3	$\bar{x}_\theta$	-2.955	-2.085	-1.311	-0.021	1.026	2.229	3.109
	s_θ	0.823	0.555	0.943	0.385	0.578	0.581	0.510
	Mni*	7.00	10.00	11.24	16.96	17.24	12.96	7.68
.4	$\bar{x}_\theta$	-3.171	-2.404	-1.346	-0.007	0.869	2.234	2.950
	s_θ	0.690	0.538	0.681	0.344	0.399	0.775	0.579
	Mni*	7.08	8.08	14.60	18.44	19.72	14.64	9.64
.5	$\bar{x}_\theta$	-3.157	-2.242	-1.051	0.160	0.941	2.340	3.117
	s_θ	0.606	0.791	0.619	0.755	0.546	0.780	0.497
	Mni*	8.16	11.4	17.04	18.64	19.28	14.32	9.48

Note. All runs made with 25 administrations per ability level, 20 item upper limit, and stepsize = .693.

*Mni = mean number of items administered

Table 10
Means and Standard Deviations
from SIM1P on ET1PL Item Pool
Varying Acceptance Range

Acceptance Range		Ability Level						
		-3	-2	-1	0	1	2	3
.1	$\bar{X}_\theta$	-2.528	-2.200	-1.174	-0.111	0.974	2.001	3.299
	S_θ	0.559	0.667	0.700	0.569	0.903	0.471	0.781
	Mni*	6.64	9.24	17.16	26.60	27.80	16.44	11.16
.2	$\bar{X}_\theta$	-2.989	-2.159	-1.144	-0.016	0.926	2.152	3.451
	S_θ	0.491	0.559	0.731	0.332	0.362	0.464	0.765
	Mni*	7.20	14.52	22.40	31.60	33.60	22.36	13.40
.3	$\bar{X}_\theta$	-3.103	-2.475	-1.162	0.003	1.016	2.161	3.024
	S_θ	0.576	0.594	0.630	0.239	0.401	0.410	0.747
	Mni*	7.60	12.40	27.96	36.72	37.88	25.32	18.16
.4	$\bar{X}_\theta$	-3.064	-2.359	-1.121	-0.094	1.043	2.073	3.054
	S_θ	0.615	0.815	0.582	0.261	0.316	0.336	0.520
	Mni*	10.20	18.00	31.40	39.00	39.36	31.52	20.12
.5	$\bar{X}_\theta$	-3.378	-2.465	-1.088	0.031	0.993	1.920	3.195
	S_θ	0.716	0.715	0.510	0.394	0.356	0.389	0.584
	Mni*	10.24	18.48	35.08	39.80	39.48	35.12	20.76

Note. All runs made with 25 administrations per ability level, 40 item upper limit, and stepsize = .693.

*Mni = mean number of items administered.

Table 11
Means and Standard Deviations from SIM1P on ET1PL Item Pool and Comparable Ideal Item Pool

		Ability Level						
		-3	-2	-1	0	1	2	3
ET1PL Pool	$\bar{X}_\theta$	-3.061	-2.175	-1.026	-0.065	1.029	1.950	2.913
	S_θ	0.561	0.460	0.573	0.469	0.516	0.696	0.533
Ideal Pool	$\bar{X}_\theta$	-3.036	-2.404	-1.037	-0.017	1.148	2.222	3.070
	S_θ	0.441	0.703	0.652	0.462	0.787	0.718	0.460

Note. All runs made with 25 administrations per ability level, 20 item upper limit, stepsize = 0.70, and acceptance range = 0.25.

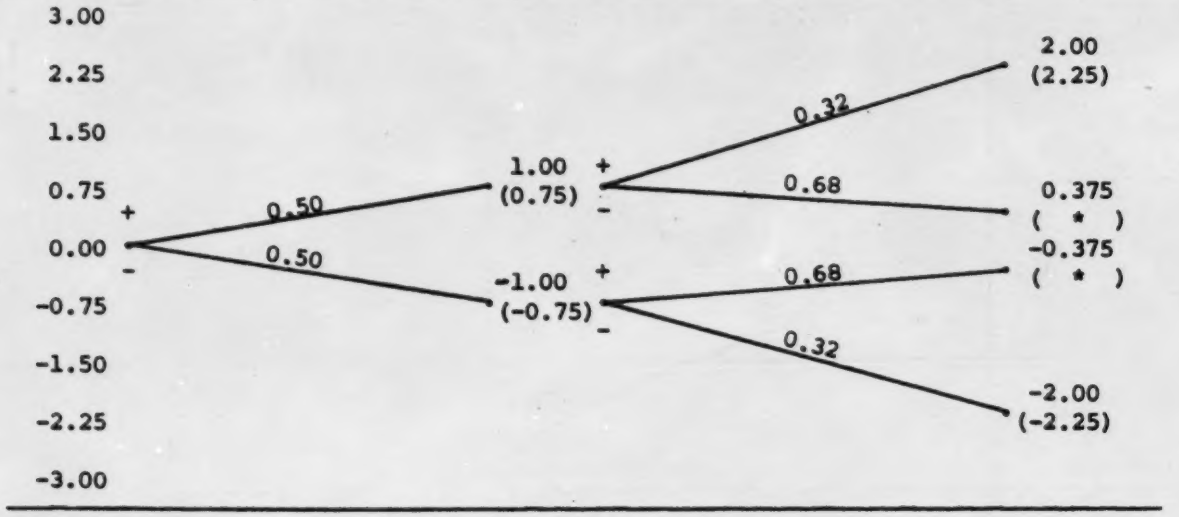
Table 12
Minimum Item Pool Requirements
for a Rectangular Idealized Pool Given
Classification Interval and Ability Range

Ability Range	Classification Interval Size	Pool Size	Minimum Session Length	Simulated Length*
(-3.5, 3.5)	0.5	15	3.9	2
(-3.5, 3.5)	0.25	29	4.86	4
(-3.5, 3.5)	0.125	55	5.8	8
(-3.5, 3.5)	0.0625	113	6.8	8
(-3.5, 3.5)	0.03125	225	7.8	7

*Note. Number of items administered to closest approximation of θ value within classification interval.

Figure 1

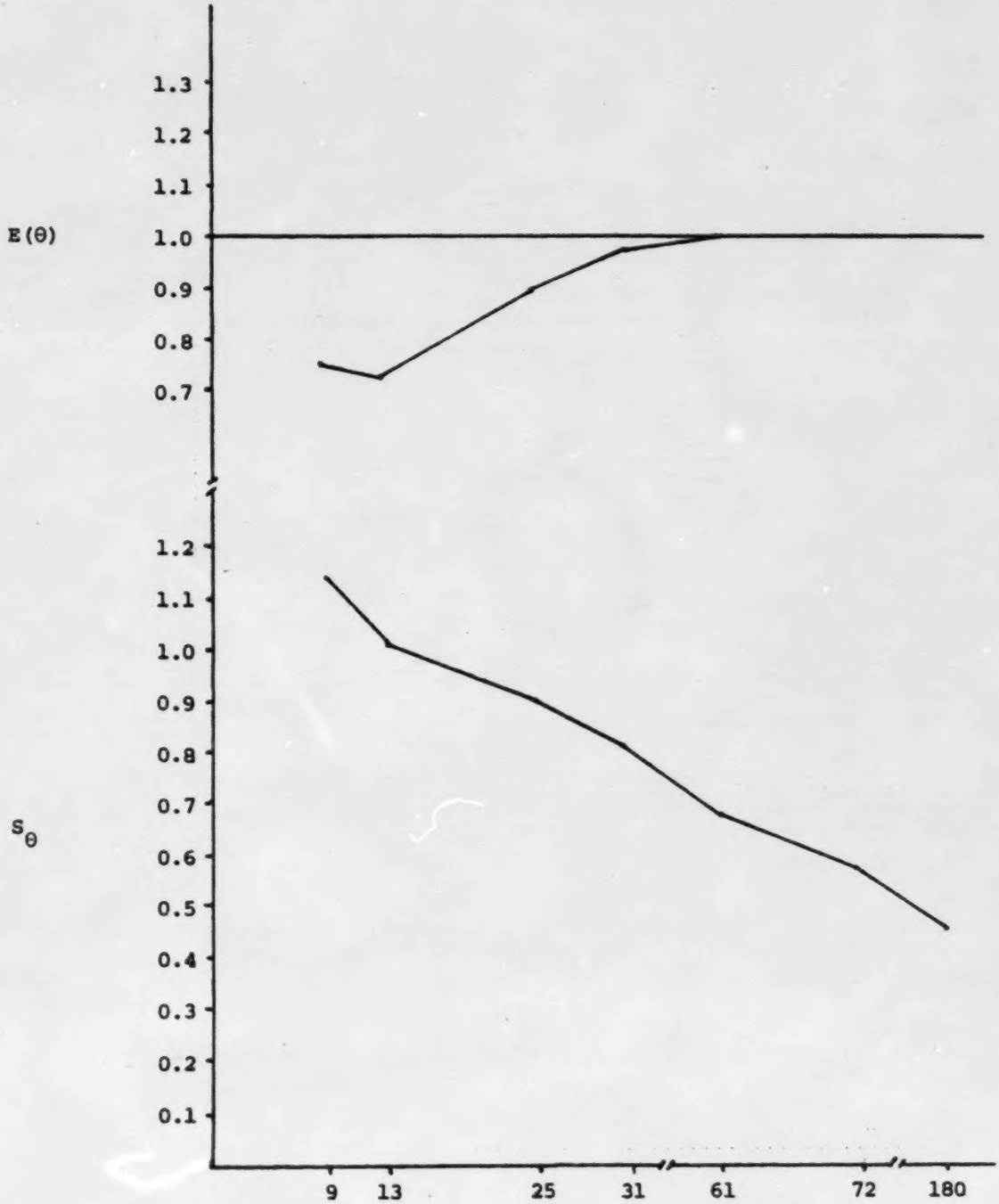
Procedural Operation of TREE1P
on a Nine Item Pool with
Stepsize = 1.0 and Acceptance Range = 0.3



$E(\theta)$	0.0	0.0
S_θ	1.0	1.174

Note. The * indicates that no item was available in the pool within $\pm$ the acceptance range.

Figure 2
Relationship Between Item Pool Size
and the E(θ) and S_θ



θ = 1.0 Stepsize = 0.693 Acceptance Range = 0.30

Figure 3
Frequency Distribution
of Difficulty Values:
72 Item VClPL Pool

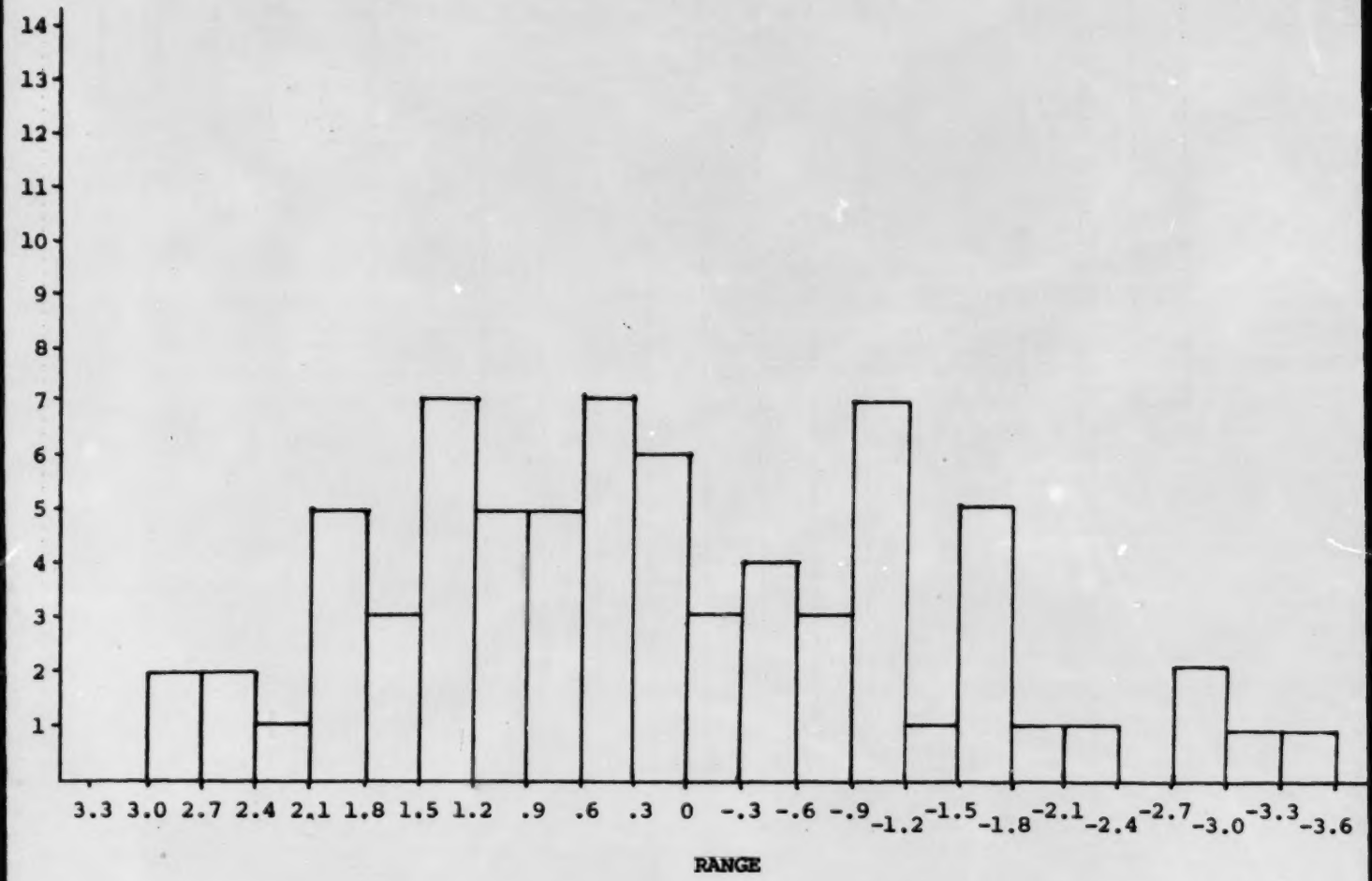


Figure 4
Frequency Distribution
of Difficulty Values:
180 Item ETIPL Pool

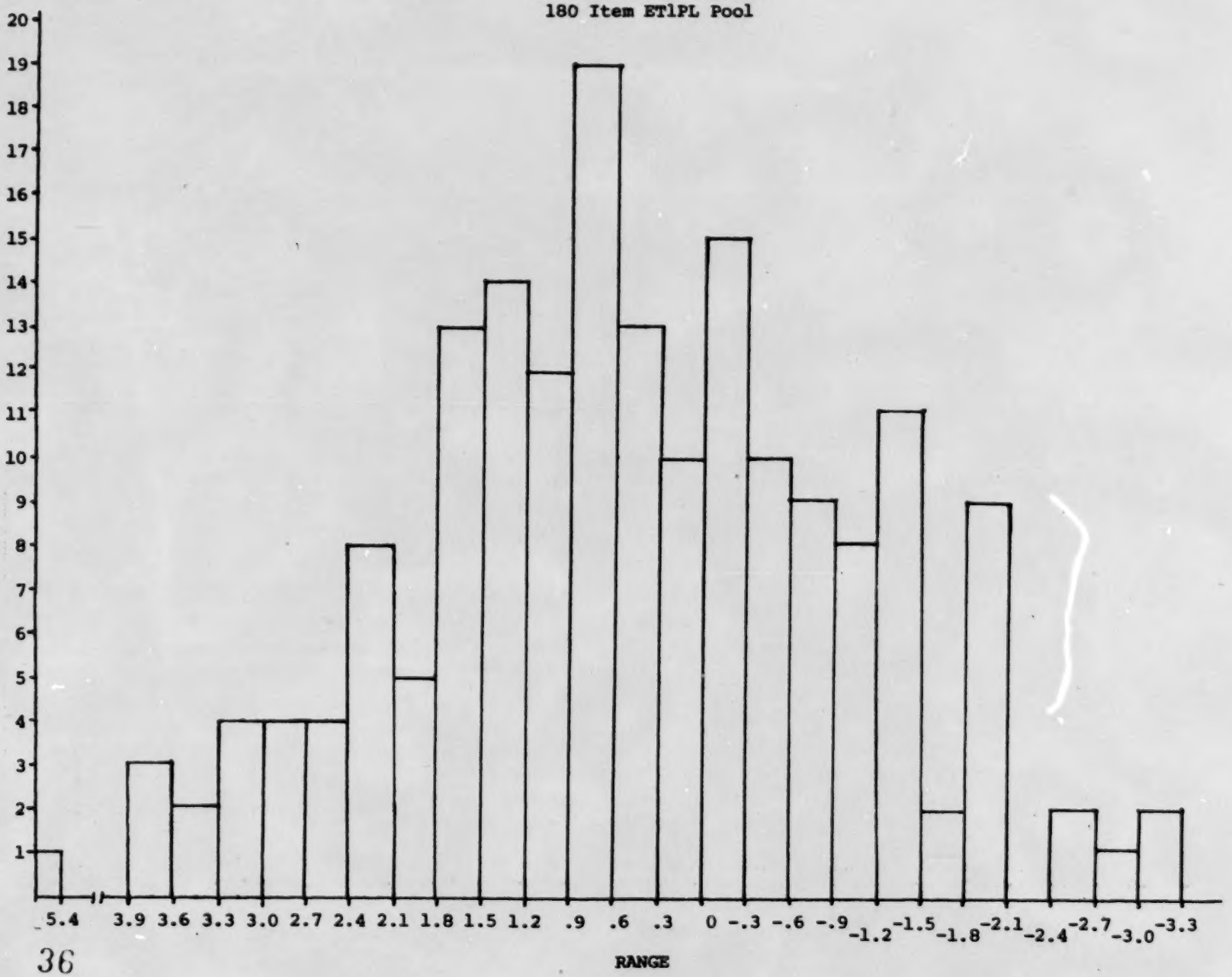


Figure 5
Relationship Between Stepsize
and the E(θ) and S_θ